

From Validation to Resilience: Sustaining Measurement Quality in AI-Assisted Enemy Item Identification

Ye Ma

Amazon Web Services

Abstract

Enemy items are item pairs that must not appear on the same test form. An automatic enemy-identification method using large language models (LLMs) has been deployed in operation. This study asks a question: how is measurement quality sustained as conditions shift after deployment? Using operational data from certification exams, two analyses examine the factors that affect the method's resilience. Analysis 1 isolates model-version and prompt updates: changing the model with the prompt held constant reduced recall from 0.75 to 0.54, while prompt refinement with the model held constant recovered it to 0.82. It also shows that most model-reviewer disagreements are edge cases and that the human standard is itself variable, with four experts spanning 0.70 to 0.91 in recall. Analysis 2 reports a disruption case: the established method works effectively on a professional-level exam but not on a foundational-level exam. A complementary content-tag based method was added to the existing method in response, with human review surfacing the disruption and validating the fix. Responsible LLM deployment in assessment requires monitoring with labeled data, testing before deployment, and evaluation systems tied to each use case, so that validity, reliability, and fairness are sustained rather than certified once.

1 Introduction

Enemy items are any two items that should not appear together on the same test form. Identifying them controls content overlap, where items measure the same construct or knowledge component; cueing effects, where one item supplies clues to another; and content overemphasis, where items disproportionately cover certain domains. Administering enemy items together threatens the reliability, validity, and generalizability of test scores, which makes enemy identification an essential task in test assembly.

Pairwise review grows combinatorially with bank size: a 500-item bank yields 124,750 unique pairs. Advances in automated item generation (AIG) accelerate bank growth, increasing manual review burden and threatening identification accuracy at the same time. An automatic enemy-identification method using LLMs was developed to address this (Ma, 2026), and has since been implemented in operation.

The deployment results bring more questions that need attentions. Deployed systems are dynamic: models change as new versions come out, banks grow as item development proceeds, and one procedure may not be able to transfer across exams with very different content. A one-time evaluation and validation cannot guarantee continued validity, reliability, and fairness. What is required instead is resilience infrastructure: the technical, procedural, and human capabilities built into the pipeline to anticipate, detect, and respond to disruption before measurement outcomes are compromised.

Therefore, this paper presents a series studies aiming to address the need of such resilience infrastructure from a responsible AI perspective. Specifically, this paper aims to identify the factors that affect the method's resilience in operation. Across deployments on multiple certification exams, we examine what destabilizes quality: prompt and model-version drift, uneven transfer across exams with differing content, the ambiguity of edge cases, and the variability of the human reviewers who serve as the comparison standard; In addition, this paper also demonstrates resilience in action through a disruption use case. The deployed method met an unexpected disruption, degraded enemy detection on an exam where the first proposed/implemented method alone was insufficient, and a complementary content-tag based method was developed in response. Human evaluation is built into this loop, surfacing the disruption and validating the fix.

2 Background

Researchers have investigated automatic enemy identification using techniques from natural language processing, machine learning, and transformer models (Becker and Kao, 2022; Meng and Li, 2024; Zhang, 2025; Brent et al, 2025; Ian et al, 2022, Mao and Zhang, 2021; Peng 2020; Weir, 2019). Becker and Kao (2022) introduced a content similarity index (CSI) approach in which a derived cutoff constrains the number of item pairs sent for manual review while maintaining sensitivity for enemy detection. Where historical data are unavailable, they propose setting the cutoff from runtime and budget constraints.

Generative LLMs offer stronger reasoning and the ability to follow complex instructions without task-specific fine-tuning, but evidence on their behavior after deployment is scarcer still. The method studied here combines both traditions: embedding-based pre-filtering to control cost, and LLM judgment to classify the prefiltered candidates.

The purpose of this paper is not to introduce a new classifier. It is an account of what happens to such a system in operation across multiple exams, model upgrades, and prompt changes, and of the infrastructure required to keep it trustworthy.

3 Method

3.1 LLM Classification with CSI-Based Pre-Filtering (Method 1)

Method 1 is the deployed method in operation. It starts with calculating item embeddings (Amazon Titan Text Embedding v2) using item texts from the stem and the key option. This combination was determined after comparing alternatives including stem with all options and stem with content-area descriptions, and Titan v2 was selected after comparing multiple embedding models. A CSI is computed as the cosine similarity of the two item vectors in each pair, and a threshold pre-filters candidates before classification. The threshold was derived based on either historical data or time and computation budget as suggested in Becker and Kao (2022).

The prefiltered pairs are then classified one pair per request by an Anthropic Claude model via Amazon Bedrock at temperature 0. The prompt supplies an exam-specific role and domain list, an enemy definition tailored to the exam, and few-shot examples of confirmed enemy and non-enemy pairs reviewed by subject-matter experts (SMEs) with

emphasis on edge cases. The model returns a label, a reason within 50 words, and a confidence rating on a 1–5 Likert scale.

The label space has two classes. Enemies are pairs whose items require the same knowledge to answer correctly, or where one item cues the other. Non-enemies are pairs whose items share only part of a knowledge component, test the same knowledge with a distinct focus or outcome, or address related subjects at different levels of specificity.

3.2 Enemy Candidate Filtering using Content Tags (Method 2)

Method 2 was developed for the disruption case, where CSI was not sensitive enough to pre-filter candidate pairs for the classifier. It replaces the similarity gate with a content partition and runs in two stages.

The first stage assigns each item a content tag naming the key knowledge the item measures via two passes. The first pass tags items one at a time, using the exam guide as context and including the tags already assigned in each prompt, so the model reuses an existing tagging system whenever one fits and the vocabulary stays compact. The second pass reviews each tag group and splits any group whose members require materially different knowledge.

The second stage detects enemies in batches rather than pair by pair. One request per tag covers every pair within that tag. A further request lists all tags and selects those pairs of tags likely to contain enemies across groups, and one request per selected pair covers those cross-tag comparisons.

4 Data and Evaluation Design

There are two sets of analyses addressing the research questions in this paper. Analysis 1 focuses on studies investigating factors that can influence the performance of method 1. Data are drawn from certification exams spanning different proficiency levels, technical complexity, and knowledge domains. Analysis 2 focuses on the disruption cases.

Analysis 1 – A series of pilot studies investigate the two study factors one at a time (Table 1). Classification metrics (recall, precision, F1 score) are used for evaluation. The two factors are model and prompt. Row 1 to 2 holds the prompt constant and changes the model. Row 2 to 3 holds the model constant and applies a major prompt update on the same exam.

Analysis 2 – Two exams were used in this anal-

Exam	Model	N pairs
Exam 1	Sonnet 3.7	113
Exam 2	Sonnet 4.0	95
Exam 2 with multiple reviewers (4 SMEs)	Sonnet 4.0	118

Table 1: Analysis 1 research design

Property	exam_ foundation	exam_ professional
Tier	Foundational	Professional
Items	642	524
Confirmed enemy pairs	431	78
Median stem length	37 words	90 words
Median key length	4 words	39 words
Items sharing a key	31.6%	0%
Derived CSI cut	0.60	0.50
Methods run	1 and 2	1 only

Table 2: Example bank summary

ysis. `exam_foundation` is a foundational-tier certification and is the disruption case; it was run with both methods. `exam_professional` is a professional-tier certification with substantially more complex content; it was run with Method 1 only and serves as the control case. Table 2 summarizes example bank situation based on items used in the analysis.

Two evaluation criteria were used: Confirmation rate is the share of flagged pairs a reviewer confirms. Because only positive predictions reach a reviewer, this is a precision quantity, not accuracy over the pair universe. Hard error rate is the share of reviewed pairs the reviewer judged not enemies at all, excluding pairs coded as edge cases with partial overlap.

5 Factors Affecting Resilience

5.1 Model-Version and Prompt Drift – Analysis 1

Changing the model while holding the prompt constant reduced recall from 0.75 to 0.54, while precision rose from 0.70 to 0.83. Changing the prompt while holding the model constant on the same exam raised recall from 0.54 to 0.82 and reduced precision from 0.83 to 0.75. This comparison shows prompt refinement recovered more recall than the model change had cost.

Later operational deployments on a newer model generation confirmed 68% of flagged pairs on Exam 3 (37 pairs) and 66% on Exam 2 (82 pairs), with 93% of the disagreed pairs coded as edge cases.

Exam	Recall	Precision
Exam 1	0.75	0.70
Exam 2	0.54	0.83
Exam 2 with multiple reviewers (4 SMEs)	0.82	0.75

Table 3: Classification results from Analysis 1

CSI threshold	exam_ foundation	exam_ professional
0.45	77.7%	100%
0.50	70.1%	96.2%
0.55	60.3%	91.0%
0.60	47.3%	76.9%
0.65	33.9%	57.7%
0.70	24.6%	50.0%

Table 4: Pre-filter recall of confirmed enemy pairs at each CSI threshold

5.2 Cross-Exam Transfer – Analysis 2

The same pipeline, code base, model, or prompt structure, performs very differently on the two example exams in Analysis 2. Table 4 reports pre-filter recall against the existing confirmed enemy pairs in the bank.

At its deployed threshold the pre-filter retrieves 96.2% of existing enemies on `exam_professional`. On `exam_foundation` the same pipeline retrieves 47.3%. Please note that using a lower CSI threshold will significantly increase the numbers of pairs for LLM processing and human review. The threshold is derived by considering those factors.

CSI pre-filtering rests on the premise that pairs with high similarity are the pairs most likely to be enemies. On a foundational exam that premise breaks in a specific, diagnosable way: two items can be genuine enemies while scoring low on CSI. One example explanation is that they converge on the same key service or the same key concept while wrapping it in different scenario or other information from the stem. The pre-filter never sees them, and no amount of downstream prompt refinement can recover them because they never reach the classifier.

The mechanism that is related to the key similarity is visible in Table 2. On `exam_foundation` the median key is 4 words against a 37-word stem; cosine similarity is dominated by scenario wording, which item writers vary deliberately. The answer space is also small and heavily reused: 31.6% of the bank sharing a key with at least one other item.

Measure	exam_ foundation	exam_ professional
Pairs reviewed	146	53
Confirmed	124	17
Confirmation rate	84.9%	32.1%
Disagreements	22	36
Edge cases	19 (86.4%)	36 (100%)
Not enemy at all	3	0
Hard error rate	2.1%	0%

Table 5: Human review outcomes for both exams in Analysis 2

Many items therefore share a key without sharing texts from stem. Key convergence is a strong enemy signal that CSI does not capture. Key similarity is one reason. There are other reasons (e.g. item length, amount of technical terms) that can lead to insensitivity from using CSI.

5.3 Edge-Case Considerations – Analysis 2

Table 5 below reports human review outcomes for both exams in Analysis 2. The two sit at opposite corners of the same trade-off.

On `exam_foundation`, 19 of 22 disagreements are edge cases with partial overlap and only 3 are outright errors, a hard error rate of 2.1%. Scoring all edge cases as correct would put confirmation at 97.9%; scoring all as errors leaves 84.9%. The defensible range is therefore 84.9% to 97.9%, and the point estimate alone overstates method error.

On `exam_professional` confirmation is far lower at 32.1%, yet all 36 disagreements are edge cases and there are no outright errors at all. The reviewer’s notes explain why: professional items share many services while pivoting on different decisions, so overlap is genuinely partial rather than absent. However, treating edge-case as enemies is controlling enemies at a more conservative way. This is also where both model and human tend to struggle the most given the subjectivity of determining such cases.

5.4 Variability of the Human Standard – Analysis 1

Because human review is ordinarily treated as the gold standard against which the method is evaluated, Analysis 1 assessed its reliability. Four SMEs independently reviewed the same pairs in the Exam 2 study (Table 6).

Reviewer recall ranges from 0.70 to 0.91 and precision from 0.71 to 0.93. The model ranks third of five on both metrics, slightly above the SME

Reviewer	Recall	Precision
SME 1	0.70	0.93
SME 2	0.88	0.71
SME 3	0.91	0.74
SME 4	0.70	0.93
SME mean	0.80	0.83
Model (Sonnet 4.0)	0.82	0.75

Table 6: Recall and precision of each SME reviewer and the model

mean on recall and below it on precision. Two reviewer styles are visible: SMEs 1 and 4 are conservative, SMEs 2 and 3 inclusive, and the model sits between them. Edge cases are where humans and model alike struggled most.

This reframes how performance should be judged. Given the variability of human review results, monitoring of consistency of human review criteria is important if using human-in-the-loop design. Additionally, using LLM-as-judge (Zheng et al., 2023) is another way to address discrepancies or variabilities from human reviewer results.

6 Discussion

The two analyses address responsible AI from two angles. Analysis 1 shows that performance changes whenever a component of the deployed workflow is updated, whether the model, the prompt, or the human standard against which both are judged, and it identifies the factors that need to be monitored or built into the evaluation system. Analysis 2 shows a disruption case in which an approved method is not sufficient on one exam as it is on others, and a complementary method was developed to address the limitation. Together they show that validity, reliability, and fairness after deployment are sustained conditions rather than properties certified once.

6.1 Model and Prompt Updates Need Re-evaluation

Analysis 1 investigated the two changes most likely to happen in a deployed pipeline: model update or prompt update. Two practices follow. First, every model-version change should trigger re-evaluation against a fixed SME-labeled set before the new version enters production. It can also evaluate whether using a more advanced model is really needed considering from a cost-benefit perspective. Second, prompt refinement should be treated as an operational practice rather than a one-time development

step, since the per-exam enemy definition and few-shot examples are what allow the model to separate merely similar pairs from true enemies.

6.2 Human Standard

One factor from Analysis 1 concerns the evaluation standard itself. Four SMEs reviewing identical pairs differed by 0.21 in recall and 0.22 in precision, and the model ranked third among the five raters on both metrics (Table 6). A method's measured performance therefore depends in part on which reviewer it is judged against. Responsible use of a human-in-the-loop design requires monitoring the reviewers as well as the model: periodic multi-reviewer calibration, written criteria for edge cases, and an LLM-as-judge pass to flag the pairs on which reviewers diverge. An LLM judge is not more accurate than an expert, but it is more repeatable, which makes it useful for locating instability in the standard rather than for resolving it.

6.3 Disruption and the Complementary Method

Analysis 2 shows a different situation. The same mechanism retrieved 96.2% of confirmed enemies at the pre-filter stage on `exam_professional` and 47.3% on `exam_foundation`. The cause was a property of the exam content. The analysis implicates three points. First, a validation study conducted on a professional-level exam would not have predicted the foundational-level result. This is particularly important when a new exam is developed. Given the exam's new feature, researchers need to be cautious of evaluating the established method on the new data. Second, the conditions under which Method 1 holds are measurable in advance. Key length, key reuse, and stem length are computable from the bank before the pipeline runs and can serve as a pre-deployment diagnostic for whether Method 1 alone is sufficient. Third, the response should match the mechanism. Lowering the CSI threshold buys recall with review volume, and the volume grows faster than the recall (Table 4). The content-tag method (Method 2) instead adds a candidate signal that captures key convergence directly. Human review results confirmed the benefits from adding this complimented content-tagging based method.

6.4 Toward Resilient Deployment

Three lessons follow for deploying automatic enemy identification, and for LLM-supported mea-

surement tasks more generally.

Monitoring with labeled data. As an automatic enemy identification system enters operation, a monitoring system should be built alongside it. The labeled data accumulated through SME review, confirmed enemies, rejected pairs, and edge-case codes, are the asset that makes this possible. Having as a high-quality evaluation set available allows to evaluate the established system whenever a model update or a prompt update is needed, so that a change in performance is observed at the time of the change rather than discovered later in production.

Testing before deployment. A pre-deployed environment is equally important, so that changes can be tested before they reach the live pipeline. This matters most when a new exam is onboarded or new item data are developed: Analysis 2 shows that a method validated on one exam can fail on another without any change to its components, and item properties such as key length and key reuse can be profiled in this environment before the method is trusted on the new bank.

Responsible AI in practice. Finally, responsible AI is not only a framework. The related principles need to be tied to each use case, with its specific requirements considered, and translated into an evaluation and monitoring system that actually enacts them. In this use case, that meant recording reason codes so that ambiguity is separable from error, estimating the variability of the human standard, and adding a complementary method when the existing one proved insufficient. A pre-deployment check certifies a system against the conditions of that moment; what sustains measurement quality as banks grow and models change is infrastructure that keeps measuring.

References

- Kirk A. Becker and Shu-Chuan Kao. 2022. Identifying enemy item pairs using natural language processing. *Journal of Applied Testing Technology*, 23:41–52.
- Ye Ma. 2026. Automatic enemy identification using large language models (LLMs) via Amazon Bedrock. Paper presented at The Association of Test Publishers (ATP) Global: Innovation in Testing.
- Xia Mao and Qiuming Zhang. 2021. An exploration of an integrated approach for enemy item identification. *International Journal of Intelligent Technologies & Applied Statistics*, 14(2):123.

- Huijuan Meng and Jinshu Li. 2024. Automatic enemy identification—are we there yet? *Artificial Intelligence and Applications*, 2(4):295–306.
- Ian Micir, Kimberly Swygert, and Jean D’Angelo. 2020. Leveraging machine learning technology to improve accuracy and efficiency of identification of enemy item pairs. *Journal of Applied Testing Technology*, pages 30–40.
- Fang Peng. 2020. *Automatic enemy item detection using natural language processing*. Ph.D. thesis, University of Illinois at Chicago.
- Brent A. Stevenor, Nadine LeBarron McBride, and Charles Anyanwu. 2025. Comparing lexical and semantic techniques for identifying enemy items via natural language processing. *Journal of Applied Testing Technology*, pages 14–20.
- John B. Weir II. 2019. *Enemy item detection using data mining methods*. Master’s thesis, The University of North Carolina at Greensboro.
- Mingqin Zhang. 2025. Enemy detection – leverage power of text embeddings. Paper presented at The Association of Test Publishers (ATP) Global: Innovation in Testing.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.